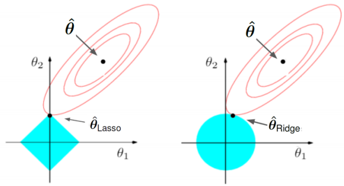
Introduction to Machine Learning

Lasso vs. ridge Regression

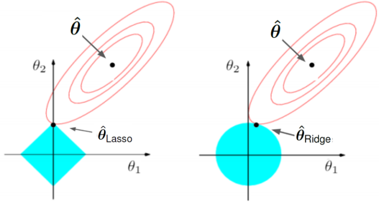


Learning goals

- Know the geometry of ridge vs. lasso regularization
- Understand the effects of the methods on model coefficients
- Understand that lasso creates sparse solutions

LASSO VS. RIDGE GEOMETRY

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 \qquad \text{s.t. } \|\boldsymbol{\theta}\|_p^p \leq t$$

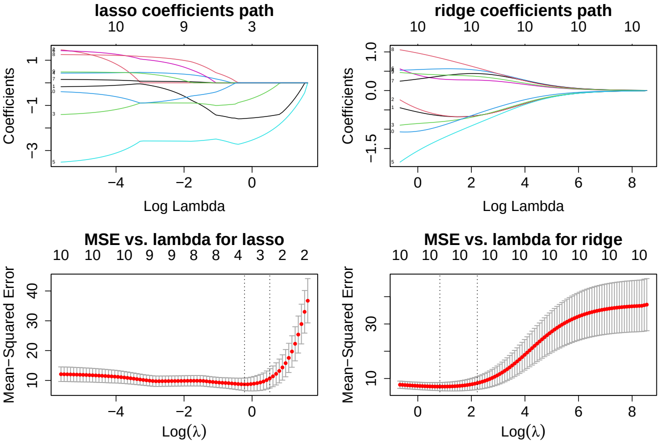


- In both cases, the solution which minimizes $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is always a point on the boundary of the feasible region (for sufficiently large λ).
- As expected, $\hat{\boldsymbol{\theta}}_{\text{lasso}}$ and $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ have smaller parameter norms than $\hat{\theta}$.
- For lasso, the solution likely touches vertices of the constraint region. This induces sparsity and is a form of variable selection.
- In the $p > n$ case, lasso selects at most n features ▸ Zou and Hastie, 2005.

COEFFICIENT PATHS AND 0-SHRINKAGE

Example 1: Motor Trend Car Roads Test (mtcars)

We see how only lasso shrinks to exactly 0.



Coef paths and cross-val. MSE for λ values for ridge and lasso.

COEFFICIENT PATHS AND 0-SHRINKAGE

Example 2: High-dimensional simulated data

We simulate a continuous, correlated dataset with 50 features, 100 observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(100)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \Sigma\right)$ and

$$y = 10 \cdot (x_1 + x_2) + 5 \cdot (x_3 + x_4) + 1 \cdot \sum_{j=5}^{14} x_j + \epsilon$$

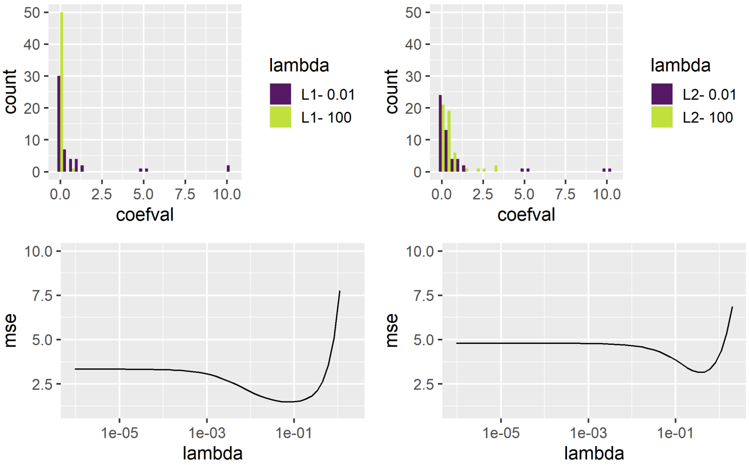
where $\epsilon \sim \mathcal{N}\left(0, 1\right)$ and $\forall k, l \in \{1, ..., 50\}$:

$$Cov(x_k, x_l) = \begin{cases} 0.7^{|k-l|} & \text{for } k \neq l \\ 1 & else \end{cases}.$$

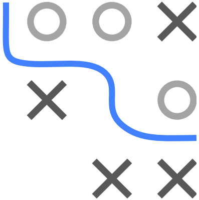
Note that 36 of the 50 features are noise variables.

COEFFICIENT PATHS AND 0-SHRINKAGE

Coefficient histograms for different λ values for ridge and lasso for simulated data along with the cross-validated MSE.



REGULARIZATION AND FEATURE SCALING



- Typically we omit θ_0 in the penalty term $J(\boldsymbol{\theta})$ so that the “infinitely” regularized model is the constant model (but this can be implementation-dependent).
- Note that unregularized LM has inductive bias of **rescaling equivariance**, i.e., if you scale some features, we can simply "anti-scale" the coefs and the risk does not change.
- Penalty methods typically not equivariant under rescaling of the inputs, so one usually standardizes the features beforehand.
- While regularized LMs exhibit low-complexity inductive bias, they lose equivariance property: if you down-scale features, coefficients have to become larger to counteract. Then they are penalized stronger in $J(\boldsymbol{\theta})$, making some features less attractive without relevant changes in data.

REGULARIZATION AND FEATURE SCALING

- Let the DGP be $y = \sum_{j=1}^{5} \theta_j x_j + \varepsilon$ for $\boldsymbol{\theta} = (1, 2, 3, 4, 5)^\top$, $\varepsilon \sim \mathcal{N}(0, 1)$
- Suppose x_5 was measured in *m* but we change the unit to *cm* ($\tilde{x}_5 = 100 \cdot x_5$):

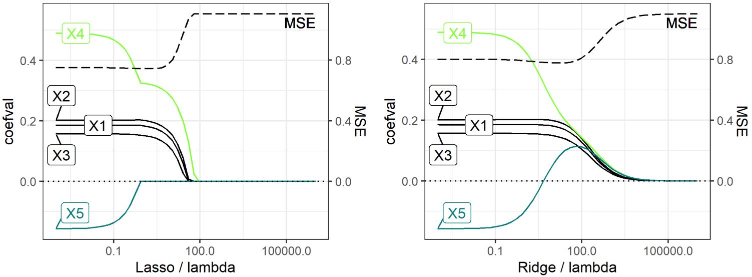
Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	MSE
OLS	0.9835872	2.147303	3.005854	3.917807	**5.20491245**	0.8124301
OLS Rescaled	0.9835872	2.147303	3.005854	3.917807	**0.05204912**	0.8124301

- Estimate $\hat{\theta}_5$ gets scaled by $1/100$ while other estimates and MSE are invariant
- Running ridge regression with $\lambda = 10$ on same data shows that rescaling of of x_5 does not result in inverse rescaling of $\hat{\theta}_5$ (everything changes!)
- This is because $\hat{\theta}_5$ now lives on small scale while *L*2 constraint stays the same. Hence remaining estimates can “afford” larger magnitudes.

Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	MSE
ridge	0.7093407	1.873643	2.661345	3.557891	**4.63642392**	1.3664731
ridge Rescaled	0.8021802	1.942568	2.675207	3.569190	**0.05134698**	1.0796400

- This also implies that for very correlated features in lasso through a unit change we could arbitrarily force a feature out of he model

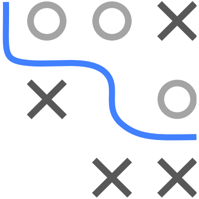
CORRELATED FEATURES



Consider $n = 100$ simulated observations using
$y = 0.2X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4 + 0.2X_5 + \epsilon$.
X_1-X_4 are independent, but X_4 and X_5 are strongly correlated.

We see that lasso shrinks the coefficient for X_5 to zero early on, while ridge assigns similar coefficients to X_4, X_5 for larger λ.

SYNOPSIS



- Neither one can be classified as overall better
- lasso can set some coefficients to zero, thus performing variable selection, while ridge regression usually leads to smaller estimated coefficients, but still dense parameter vectors $\boldsymbol{\theta}$.
- Lasso is likely better if true underlying structure is sparse, so if only few features influence y. Ridge works well if there are many (weakly) influential features.
- Lasso has difficulties handling correlated predictors. For high correlation ridge dominates lasso in performance.
- For lasso one of the correlated predictors will have a larger coefficient, while the rest are (nearly) zeroed. The respective feature is, however, selected randomly.
- For ridge the coefficients of correlated features are similar.
- For references, see ▸ Tibshirani, 1996 and ▸ Zou and Hastie, 2005